DSP BASED ACOUSTIC VEHICLE CLASSIFICATION FOR MULTI-SENSOR REAL-TIME TRAFFIC SURVEILLANCE

Andreas Klausner, Stefan Erb, Allan Tengg, Bernhard Rinner

Graz University of Technology
Institute for Technical Informatics
Inffeldgasse 16/1, Austria, Graz, 8010
{klausner, erb, tengg, rinner}@iti.tugraz.at

ABSTRACT

Vehicles may be recognized from the sound they emit when driving along a road. Characteristic acoustic finger prints and audio features can be used to increase the robustness of existing video based vehicle tracking and classification algorithms. Using this information in a multi-sensor surveillance system helps to improve various parameters such as recognition rates, detection times and robustness. We propose a two-fold approach, where vehicle detection and classification are handled separately. We demonstrate the feasibility of the proposed method using outdoor audio sequences of traffic situations.

1. INTRODUCTION

In the I-SENSE project [1, 2] we develop an intelligent multi-sensor fusion framework for embedded online data fusion. Fusing data from various sensors helps to improve the robustness and confidence, to extend the spatial and temporal coverage as well as to reduce ambiguity and uncertainty of the processed sensor data. In the I-SENSE project we exploit these characteristics to improve the quality of traffic surveillance.

Since current traffic surveillance systems (e.g., *SmartCam* [3]) are primarily based on video, integration of data from audio, infrared, supersonic and inductive loop sensors helps to improve various parameters such as recognition rates, detection times, robustness and quality of service. While acoustic surveillance systems have been well studied (e.g., recognition of vehicles [4, 5], machines and dropping objects [6]), multi-sensor data fusion approaches are currently driven by automatic speech and gesture recognition systems [7].

Almost all vehicles emit characteristic sounds when moving on a road. The sound is mainly composed of (i) rotational parts and vibrations in the engine (ii), noise caused by the exhaust tube (iii), friction between the tires and the pavement and (iv) broad band noise caused by the air stream of moving vehicles. In this article we describe our ongoing research on robust acoustic feature extraction methods to support real-time traffic surveillance. Currently, recorded traffic sounds are analyzed with respect to four criterions: (i) the presence of a vehicle (ii), the characteristic acoustic fingerprint of a vehicle used to track an object (iii), the average velocity together with the driving direction, and (iv) the vehicle category a detected object belongs to. Our acoustic classification system is designed for distinction between three different vehicle categories car, van and truck. In our recording data only these classes appear without presence of other vehicle categories such as motorcycles, but the system can easily be adapted for distinction of further categories.

The reminder of the paper is organized as follows: Section 2 presents the utilized experimental setup and microphone configuration for our research. Section 3 discusses our approach for vehicle detection while section 4 focuses on acoustic vehicle tracking. In section 5 different feature generation algorithms are presented, to extract a set of audio features from the input data. In section 6 it is utilized for distinction between the three vehicle categories together with a Support Vector Machine (SVM) classifier. Section 7 presents the experimental results and shows the feasibility of our approach. Section 8 concludes the paper with a summary and an outlook on future work.

2. ACOUSTIC TRAFFIC SURVEILLANCE SETUP

The setup for our acoustic traffic surveillance consists of two microphones next to the road. The distance between the sensors (microphone base) is set to 1 *m* in order to permit a cross correlation analysis (see section 5, and the height above ground is 1 *m*. Traffic sounds have been recorded at a sample frequency $f_S = 8\,kHz$ in 16 *bit* resolution together with video data in order to ease the evaluation.

These recorded traffic sounds have then been utilized for the development of the vehicle detection and classification methods in MATLAB. For real-time detection and classification these algorithms have been ported to and optimized for a TMS320DM642 signal processor from Texas Instruments.

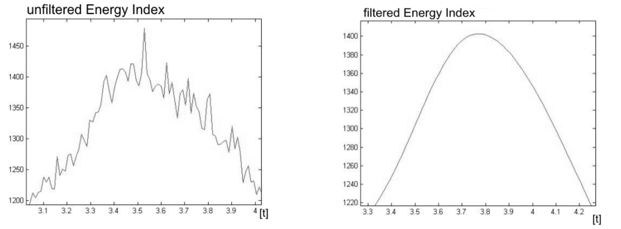
3. EVENT AND VEHICLE DETECTION

Our approach for vehicle detection is two-fold. To keep the required computation resources low, we have decided to separate the detection of vehicles from their classification. For the vehicle detection we use a simple and fast algorithm as presented in the following. If this simple algorithm determines an interesting acoustic event, the complex and time consuming classification algorithm is triggered for an in depth analysis of the observed object.

In order to find out, if a vehicle is in range of the microphone pair, an index is needed that describes the energy density of the input signal as function of time. A pure energy analysis in time domain is not suitable for that, since the index must be particularly robust against background noise from the (e.g., noise caused by the wind) environment. In addition the method should offer a quantity for the probability of a valid vehicle passing. In our approach we group the input samples into hamming windowed blocks, apply a short-time FFT analysis and sum up the logarithm of the spectral line amplitudes:

$$E[i] = \sum_k log\left(\mathbf{X}_i[k]\right) \qquad (1)$$

where $\mathbf{X}_i[k]$ is the Fourier transform of the i^{th} input block $x_i[n]$ and E_i denotes the corresponding energy index. A good tradeoff between required processing power and accuracy of the energy index is given by a blocksize $N = 256$ samples. The hopsize between adjacent blocks depends on the actual setup and expected disturbances. The index is shown during a vehicle passing in figure 1. In order to detect a vehicle, it is necessary to find significant maxima in this energy function. Applying a smoothing filter eases the implementation of an algorithm which is capable of finding local maxima. A butterworth structure best smoothes the energy course.



(a) according to equation 1 (b) filtered with Butterworth

Figure 1: Energy index, E_i during same vehicle passing by: (a) unfiltered and (b) filtered

On the first view (cp. figure 1) the group delay t_d (in our case of $8kHz$, $t_d = 0.29s$) of the butterworth filter may look problematic. But this delay is relatively small compared to the peak detectors retention, described in the following.

There are many events that result in a peak in the energy index. So it is necessary to extract discriminating features to decide whether a peak is caused by an object of interest (e.g., car, truck). We have identified the following five conditions for a robust vehicle detection. The most important parameter of a peak is the amplitude. Peaks with an amplitude lower than a specified threshold, $Thres$ (cp. fig. 2, (1)) are ignored. They must have a duration in a certain time range $WdMx$ (cp. fig. 2, (4)). Furthermore, the raise-time (cp. fig. 2, from (2) to (1)) and the fall-time (cp. fig. 2, from (1) to (3)) must be lower than a predefined value $Diff$. Vehicles typically cause a symmetric peak, which distinguishes them from disturbances. Parameter $DiffMM$ is defined as the difference between the left and the right minimum in the time window. The last criterion that must be fulfilled is the area below the energy graph in the specified time window. It must exceed the parameter $MinArea$. Trucks with trailers typically create peaks with more than one local maximum. Therefore, a time span parameter $LocMxWd$ is introduced to prevent multiple detections caused by a single vehicle.

4. VEHICLE TRACKING

In contrast to vehicle detection, vehicle tracking based on acoustic fingerprinting requires substantial computation. In our approach a fingerprint is generated if a vehicle passing the microphones has been detected. Problems, which have to be considered, especially when dealing with vehicles at higher speed are (i) the increased noise caused by the tires and air disturbances, (ii) the reduced time slot for vehicle recording and thus, less accurate fingerprint information and (iii) the Doppler effect. All three influences always occur together and significantly complicate extraction of reliable fingerprints. Distinct harmonic peaks from the motor sound can be recognized from individual FFT spectra only in rare cases.

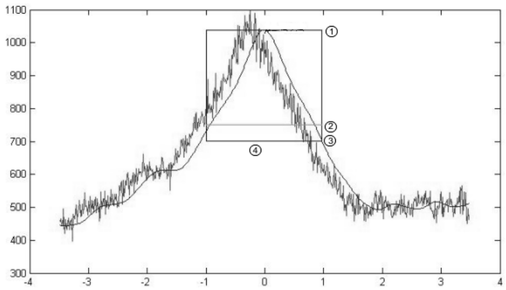


Figure 2: Energy index $E[i]$ with conditions for vehicle detection

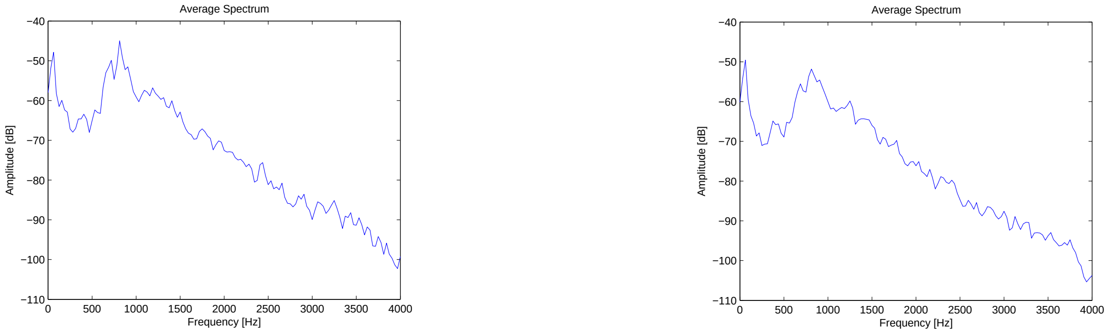
Symbols	Description
$Thres$	Threshold for peak
$WdMx$	Window size
$Diff$	Difference of min. and max.
$MinArea$	Minimal area of valid peaks
$DiffMM$	Measure for symmetry
$LocMxWd$	Suppressor of peak ripple

Table 1: Parameters for our experimental setup

Our so called audio fingerprints are calculated from characteristic FFT spectra of vehicles. In order to obtain more precise information about spectral composition of engine noise and broad band characteristics, an averaging of many FFT spectra is obvious. Figure 3 visualizes an example of an averaged spectrum as mentioned above. The time span, which is taken into consideration for a precise identification of a vehicle, should be as large as possible. For building the average spectra, the time window is centered around the detected energy peak. For the implemented algorithm, a time period of approximately $1.5s$ around the maximum has been chosen. To avoid a loss of information the input signal is windowed by 50% overlapping Hann windows. A longer time period has to be avoided for the audio fingerprint, since in case of a high traffic density there might be a temporal overlap with adjacent vehicles.

The Doppler effect smears the spectrum of passing vehicles proportional with increasing velocity, which theoretically means, that a correct evaluation of FFT spectra can only be carried out during their approach.

In order to track a vehicle over multiple sensors, it is required to recognize it on other audio channels. Within small spatial ranges an acoustic fingerprint can be used for this recognition and is an easy objective for an algorithm by using the average spectra to compute a numeric value which describes the similarity of two spectra. This similarity measure is calculated by cumulative summation of the weighted difference between each frequency bin when comparing two acoustic fingerprints. The weighting coefficients w_k decrease



(a) Vehicle passing station A

(b) Vehicle at station B

Figure 3: Averaged spectrum of vehicle (50 kph) recorded with 50 m spatial difference

proportional to the frequency bin index k with $w_k = \frac{1}{k}$. In this way, lower frequencies from the motor noise of vehicles are more weighted.

5. VEHICLE FEATURE EXTRACTION

Various signal processing algorithms were implemented with MATLAB in order to collect a pool of candidate features able to distinguish between our three vehicle categories. Each of the algorithms extracts several features from the raw input data and returns so called candidate features. They are used as input to an optimization stage of the system design, where the subset with best class discrimination ability is selected out of the candidates. This optimization procedure is performed with a genetic algorithm (GA) [8] that utilizes the classification performance (percentage of correctly classified vehicles) of the classifier from section 6 as quality measure for class discrimination properties of a selected feature subset. The goal of the optimization is to find the feature subset with best classification properties. Extracted candidate features are calculated with the algorithms described in the following.

5.1 Time Domain Features

Features in time domain are generated from short time energies, zero crossing rates and correlation analysis algorithms. Due to block processing of the audio signals, spectral features are always given as feature vectors which reflect signal behavior over time. Thus, statistical moments such as mean, variance and median values must always be utilized to reduce feature data and to include information about nonstationary feature behavior. In speech recognition short time energy is used to discriminate between voiced and unvoiced speech signals. For acoustic vehicle classification the mean energy within the analysis window is an important feature, as large trucks usually produce much more noise than other vehicles. The crucial step is to find a suitable length for the analyzed signal. If the analysis window is chosen too long, in dense traffic situations adjacent vehicles appear in the energy course. Conversely if it is too short, non-stationary signal behavior and noise effects may scatter feature data. Thus, the analysis time interval is selected depending on on the present traffic scenario with expected disturbances and vehicle velocities. The zero crossing rate counts the number of zero crossings of a signal within the specified time interval. It is a measure for the noisiness of a signal. As cars produce more tire noise than trucks at higher frequencies (typical frequency range 500Hz-2kHz), this measure provides useful class discrimination properties, especially with higher vehicle velocity. Cross-correlation analysis can be performed with our two microphones placed along the road side. Point like sound sources produce interference patterns in a two dimensional diagram, where the cross-correlation function is plotted over time. By applying image processing algorithms, vehicle characteristic information can be extracted from these traces: The speed can be estimated, the number of axles and their spacing.

5.2 Spectral Features

Spectral features include signal attributes that describe average energies, positions and spreads in frequency domain, such as the spectral centroid, signal bandwidth, spectral flux, or band energy ratios. Mathematical definitions can be found in [9]. They are commonly used in speech recognition, environmental sound recognition and audio genre classification, and provide feature candidates with useful information about spectral signal properties. Because single spectral bins do not contain relevant information for classification purposes, and are also mutually correlated (i.e. they are linearly dependent on each other), spectral bins that provide good classification performance achieve only little performance improve when combined together in a feature vector. As single feature values provide only local information for distinct blocks, again statistical moments must be calculated to capture long term signal characteristics from the analysis window of a passing vehicle.

5.3 Cepstral Features

The Cepstrum $c(\tau)$ of a signal $x(t)$ is defined as the inverse Fourier transform of the logarithm of its spectrum:

$$c(\tau) = F^{-1}\{\log|F\{x(t)\}|\}, \tag{2}$$

where F denotes the Fourier transform. Cepstral coefficients (CCs) are popular feature candidates in speech recognition systems, as they provide very good information packing properties: Low order CCs capture information about the slowly varying properties of the spectrum, also referred to as spectral envelope. Multiplication of the signal by a constant gain for example, only affects the first cepstral coefficient (c_0 term), feature vectors can thus be made invariant to changes of gain by exclusion of this term. Higher order cepstral coefficients can also be used to detect the fundamental frequency of a periodic signal, because harmonic line sets in the logarithmic spectrum coincide as single peaks in the cepstral domain.

CCs are computed either directly using equation 2, or estimated via linear predictive analysis by converting LPC coefficients into LP based cepstral coefficients. The LPC parameters a_i in an autoregressive (AR) model are directly obtained as system of equations from the autocorrelation function $r(k)$, by solving the so called Yule-Walker equations:

$$\sum_{i=1}^{p} a_i r\left(|k-i|\right) = r(k), \tag{3}$$

where p denotes the selected model order, which must be set high enough to provide a detailed signal information. The set of linear prediction coefficients (LPCs) a_i is converted into LP based cepstral coefficients by the following recursion:

$$c_m = \begin{cases} \ln\left(r(0)\right) & \text{for } m = 0 \\ a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right) c_k a_{m-k} & \text{for } m \geq 1 \end{cases} \tag{4}$$

where $a_0 = 1$ and $a_k = 0$ for $k > p$.

This method avoids any signal transformation and thus, offers highly reduced computational effort, provided that only a few cepstral coefficients are needed – which is the case. LP based CCs as features afford efficient use in real time environments. In our traffic scenario case study both FFT and LP based CCs proved good classification results and outperformed other spectral envelope estimates, such as filter bank analysis (Haar transform, channel vocoder) and direct utilization of LP parameters a_i. Figure 4 shows the class separation based only on c_0.

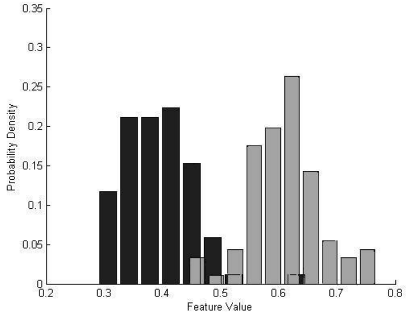


Figure 4: Histograms for feature values of c_0. Car (left) and truck (right) classes are almost completely separated.

6. VEHICLE CLASSIFICATION

The vehicle classification is performed with a modified support vector machine (SVM) [10, 11] classifier which is trained with a given amount of vehicle samples before switched to decision mode. In order to evaluate classification performance, a vehicle database provides 200 vehicle samples per class. An important goal for our acoustic classification system is to find features able to keep class discriminative capabilities when utilized in different traffic scenarios. Therefore the present database combines vehicles recorded on both urban roads and suburban highways, and thus, vehicles moving with low and high speed in a range from 30kph to 100kph. During the optimized feature subset search this let to features which achieve reliable classification performance for both traffic situations. Hence, they are generally able to discriminate between the vehicle classes without influence of different velocities. This technique is also referred to as generalization of a feature subset.

Scenario	**Vehicles**	**dLC**	**dRC**	**FP-ok**	**FP-nok**	**Rate**
1	33	32	31	31	0	93.9%
2	36	28	32	27	5	75.0%
3	38	36	37	23	14	60.5%
4	56	49	46	31	18	55.4%

Table 2: Experimental results object detection

7. EXPERIMENTAL EVALUATION

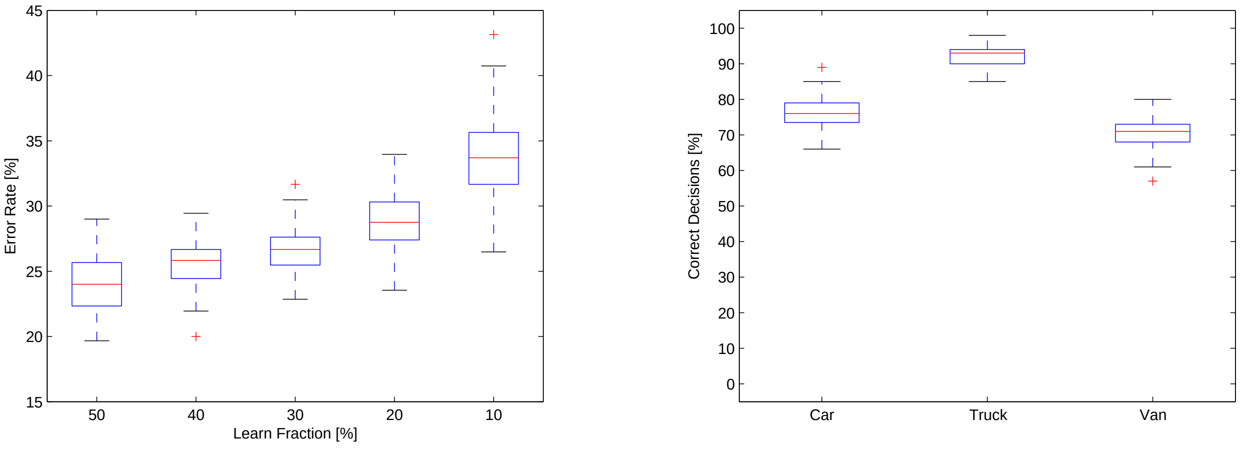
Different algorithms have been developed for the real-time analysis of traffic sounds at the DSP. The vehicle detection is a C-implementation of the algorithm described in section 3. The highly optimized code for data acquisition and the event detection causes a utilization of approximately 10% of the DSP. The identification assigns a characteristic fingerprint which is compared with all stored fingerprints of a certain time span on the other channel. In case of matching finger prints the event is counted and the average velocity is estimated. Furthermore, the driving direction of vehicles can be determined.

In the following we demonstrate the feasibility of our approach based on four different test scenarios:

Scenario 1: urban one-way street (max. 50 kph)
Scenario 2: urban street, two driving directions
Scenario 3: suburban two lane street (max. 70kph)
Scenario 4: highway (max. 100kph)

The second column of table 2 contains the quantity of vehicles during the scenario. **dLC** and **dRC** comprise the detected events at the left and the right channel. The following two columns (**FP-ok**, **FP-nok**) present the correct as well as the incorrect matches of the implemented fingerprint algorithm.

A major problem when evaluating our classification system is the scattering of the results, when using randomly selected training data sets. Since vehicle features may contain strong outliers, the classifier can easily be confused when trained with noised learning samples. Hence, classification error rate highly depends on the utilized learning data. A solution to this problem is to train and test the SVM several times, each with different learning samples chosen randomly from the database and to present the resulting data scattering as shown in figures 5. This procedure leads to more accurate evaluation results, as the spread of classification performance is estimated. In figure 5a performance evaluation of the optimized feature subset is carried out by exploring the error rate when trained with different percentages of the database size, here referred to as learning fraction. If the database contains 200 vehicles from each class, and the SVM for example is trained at 40% learning fraction, this corresponds to 80 vehi-



(a)

(b)

Figure 5: Error rate evaluation for (a) different learning data size, and (b) classifier performance at 50% learning fraction.

cles from each of the categories used for learning and the rest for evaluation. In order to show the scattering of error rate, evaluation is performed 100 times with randomly selected learning data. Figure 5b contains the percentage of correct decisions in each of the classes achieved at 50% learning fraction. As we can see, the truck class is well distinguishable, while car and van classes can't be fully separated from each other.

8. CONCLUSION

As shown in the table 2 the implemented algorithm for event detection (vehicle detection) is suitable in test cases with an average velocity below 70 *kph*. Increased background noise (e.g., caused by wind) makes proper vehicle detection more complicated. The fingerprint algorithm works best for slow moving vehicles and one-way driving direction. If the characteristic sounds of vehicles overlap, as it is the case on two-lane streets, the number of identified vehicles drops. According to the results shown in figure 5b, acoustic vehicle classification based on acoustic-features only is not that reliable, but it is a well suited extension to other sensory data (e.g., video, inductive loops).

REFERENCES

[1] A. Klausner, B. Rinner, and A. Tengg, "I-SENSE: Intelligent embedded multi-sensor fusion," in *Proceedings of the 4th IEEE International Workshop on Intelligent Solutions in Embedded Systems (WISES)*, Vienna, Austria, June 2006, pp. 105–116.

[2] A. Tengg, A. Klausner, and B. Rinner, "I-SENSE: A Light-Weight Middleware for Embedded Multi-Sensor Data-Fusion," in *Proceedings of the 5th IEEE International Workshop on Intelligent Solutions in Embedded Systems (WISES)*, Madrid, Spain, June 2007.

[3] Michael Bramberger, Andreas Doblander, Arnold Maier, Bernhard Rinner, and Helmut Schwabach, "Distributed Embedded Smart Cameras for Surveillance Applications," *Computer*, vol. 39, no. 2, pp. 68–75, Feb. 2006.

[4] E.M. Munich, "Bayesian subspace methods for acoustic signature recognition of vehicles," in *Proceedings of the European Signal Processing Conference (EUSIPCO-04)*, Vienna, Austria, Sept. 2004.

[5] B. G. Ferguson and K. W. Lo, "Extracting tactical information from acoustic signals," in *Proceedings of the International Conference on Intelligent Sensors, Sensor Networks and Information Processing ISSNIP*, Melbourne, Australia, Dec. 2004.

[6] A. Harma, M.F McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *Proceedings of the International Conference on Multimedia and Expo*, Amsterdam, Netherlands, Jul. 2005.

[7] M.N. Kaynak, Zhi Qi, A.D. Cheok, K. Sengupta, and Ko Chi Chung, "Audio-visual modeling for bimodal speech recognition," in *Proceedings of the International Conference on Systems, Man, and Cybernetics*, Tucson, USA, Oct. 2001, pp. 181–186.

[8] M. Mitchell, *An introduction to genetic algorithms*, MIT Press, Cambridge, MA, 1996.

[9] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. on Speech and Audio Processing*, vol. 7, pp. 293–302, 2002.

[10] A. Klausner, A. Tengg, and B. Rinner, "Enhanced Least Squares Support Vector Machines for Decision Modeling in a Multi-Sensor Fusion Framework," in *Proceedings of the International Conference on Artificial Intelligence and Pattern Recognition (AIPR-07)*, Orlando, US, July 2007.

[11] A. Klausner, A. Tengg, C. Leistner, S. Erb, and B. Rinner, "An audio-visual sensor fusion approach for feature based vehicle identification," in *Proceedings of the International Conference on Advanced Video and Signal based Surveillance (AVSS-07)*, London, GB, September 2007.